META-ANALYSIS OF LIVER TRANSCRIPTOMIC DATA IDENTIFIES ACCURATE DISEASE CLASSIFIERS AND DISEASE PERTURBED NETWORKS

BY

YULIANG WANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Chemical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Advisor:

Professor Nathan Price

ABSTRACT

Chronic liver diseases are a major health problem. Previous DNA microarray studies of different liver diseases have improved our knowledge of the molecular pathogenesis of liver diseases and produced potential biomarkers. However, these studies typically rely on binary phenotype comparisons (e.g. cancer vs. normal) to identify disease signatures. It is possible that the resulting signatures may be partially shared by other liver diseases not included in the binary comparison. In this study, we took a comprehensive and organ-specific approach, where we studied all liver pathophysiological states in a single unifying context, and found a specific transcriptomic signature for each phenotype with respect to all the other phenotypes, instead of just one. The resulting 36-gene disease signature had 85% accuracy in 10 fold cross validation. Through stringent leave-one-lab out independent validation, we found that high classification accuracy was achieved when there was a total of around 100 samples from 2 independent contributing labs. We also identified perturbed networks in liver diseases in general and hepatocellular carcinoma in particular. Many of the classifier genes and perturbed networks are involved in important biological processes in liver disease pathogenesis, including immune response and inflammation, fibrogenesis, metabolism and its regulation, apoptosis, and cellular signaling. The disease classifiers and perturbed networks identified in this study may be potential candidates for novel diagnostic approaches to multiple liver diseases.

ACKNOWLEDGEMENTS

The author wishes to express sincere appreciation to Professor Nathan Price for his assistance in the preparation of this manuscript and for inspiring discussions in this thesis project. In addition, special thanks to Jaeyun Sung for helping me to initiate this project, and for informative ideas and discussions along the way. Thanks also to Andrew Magis who helped with data processing and to Shuyi Ma for helpful discussions. And finally, thanks to my family and girlfriend who have helped me through the process of earning my Master's degree.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION 1
CHAPTER 2. RESULTS 3
2.1. Microarray data collection and rigorous preprocessing 3
2.2. An overview of the data set 3
2.3. Overview of a coarse-to-fine search algorithm for liver disease classification 4
2.4. Overview of method to identify deregulated networks in liver pathogenesis 4
2.5. Top-Scoring-Pair based decision tree performs highly accurate classification 4
2.6. Classification performance in independent validation 5
2.7. Classifier genes are closely involved in liver disease pathogenesis 6
2.8. Perturbed pathways in liver diseases 8
2.9. Tables 10
2.10. Figures 14
CHAPTER 3. CONCLUSION 20
CHAPTER 4. METHODS 22
4.1. Building a coarse-to-fine diagnostic tree and identifying branch-based classifiers 22
4.2. Identifying node-based classifiers 22
4.3. Classification of microarray samples 23
4.4. Identifying perturbed networks 23
REFERENCES 25

CHAPTER 1

INTRODUCTION

Liver diseases pose a major health threat worldwide. The prevalence of hepatitis C infection is approximately 2.2-3.0% (130–170 million people) worldwide [1]. Besides high prevalence, hepatitis C shows a higher propensity to yield chronic infection: HCV establishes chronic infection in 60-80% of infected individuals [2], making it a major risk factor for chronic liver diseases, especially in western countries and Japan. It is estimated that 1% of populations have histological cirrhosis, most of them undiagnosed [3]. 80% cases of HCC develop in cirrhotic livers, and cirrhosis is the strongest predisposing factor [4]. Hepatocellular carcinoma (HCC) is the fifth most common cancer worldwide and the third most common cause of cancer mortality [5]. It is estimated that there were 564,000 new cases worldwide and almost as many deaths in 2000 [6].

Previous studies of liver transcriptome using DNA microarrays have advanced our understanding of liver diseases and yielded potential diagnostic markers [7]. However, these studies rely on binary phenotype comparisons (e.g. normal liver vs. cirrhosis, normal liver vs. HCC, cirrhosis vs. HCC) to identify disease signatures. It is possible that the resulting signatures may be partially shared by other liver diseases not included in each binary comparison. Therefore it is important to take a comprehensive and organ-specific approach, where we study all liver pathophysiological states that meet our phenotype and data inclusion criteria in a single unifying context, and find a specific transcriptomic signature for a phenotype with respect to all the other phenotypes, instead of just one.

Recently, several classification methods that use the relative expression ordering of only a few genes have been developed [8-10]. The decision rules of these methods usually follow this formalism: if a particular gene *i* is expressed higher than another gene *j*, a sample is classified as class 1; if not, the sample is classified as class 2. These methods show comparable accuracy with conventional methods such as support vector machines (SVM) while use very simple decision rules that involve small number of genes. Since these methods rely only on gene expression ranks, not on the absolute intensities, they are invariant to data normalization methods and are free of parameter fitting [8].

These methods have been extended to deal with multi-class classification problems by integrating with a coarse-to-fine search algorithm [11]. In this study, we did a rigorous meta-analysis of available liver transcriptomic data to identify disease-specific transcriptomic signatures using this relative-expression based method. The disease signatures we identified have 85% classification accuracy in cross validation yet involve very few genes (36 unique features).

As DNA microarrays become widely used, there are many meta-analysis studies that combine data from multiple studies to identify robust disease signatures [12]. In a meta-analysis where disease classification is the end-point, independent validation is crucial. One previous study found that without any data

integration, training classifiers on one data set and testing it on the other independent data set resulted in poor classification accuracy [13]. Given the wide-spread differences in sample characteristics, sample preparation, hybridization, and other protocol differences between labs, this is not unexpected [14]. In our meta-analysis, we did a series of stringent independent validations and found that as long as there are around a total of 100 samples from two independent labs to train our classification method, highly accurate classification can be achieved on a completely independent data set without any data integration. This shows that by combining data from multiple labs in a rigorous way, batch effects would be mitigated, robust disease signatures can be identified, and high classification accuracy can be achieved on future independent test data.

Besides accurate disease classifiers, we also identified perturbed networks in liver disease pathogenesis. We did two relevant comparisons: one is to compare all liver diseases included in this study (chronic hepatitis C, cirrhosis, HCC) with normal liver in order to identify commonly perturbed pathways during liver disease pathogenesis in general; the other is to compare HCC with all other non-malignant liver phenotypes (normal liver, chronic hepatitis C, cirrhosis) in order to identify perturbed pathways in hepatocellular carcinogenesis. We used differential rank conservation (DIRAC) [15], which is also based on the relative expression of genes within networks, for this analysis. Many of the classifier genes and perturbed networks are involved in important biological processes in liver disease pathogenesis, including immune response and inflammation, fibrogenesis, metabolism and its regulation, apoptosis, and cellular signaling.

CHAPTER 2

RESULTS

2.1. Microarray data collection and rigorous preprocessing

We first established our phenotype and data set inclusion criteria. We included phenotypes that have at least 20 samples in total and at least two independent contributing labs. All studies should be based on Affymetrix microarray platforms to ensure measurement consistency and facilitate data integration. All samples should have well-labeled phenotypes. After applying these criteria, we collected 392 microarray samples from 7 different studies [16-22]. This data set consists of normal liver, chronic hepatitis C (CHC), cirrhosis and HCC (Table 1). We used disease samples of HCV etiology (i.e. they all have persistent HCV infection) alone, since samples of varying etiologies would obscures the data and limit the pathogenetic insights gained from transcriptomic profiling [2, 23]. We downloaded the raw intensity files of these 392 samples from Gene Expression Omnibus and used our own microarray consensus preprocessing pipeline to process the raw data. This pipeline went through all three different Affymetrix platforms (U133A, U133 2.0 and U133 plus 2.0), found common probe sets across platforms, built a consensus platform, used the Matlab (Mathworks, Inc) implementation of GCRMA [24] to preprocess all samples together, and used MAS5 detection algorithm [25] to make detection calls (present, absent, marginal). After this preprocessing step, we have 22277 probes. We downloaded the most recent Affymetrix official probe annotation file (as of April 1, 2011), and removed probes that do not map to any known genes according to the latest annotation. After this step, we had 20928 probes. To ensure that a probe is reliably detected, we decided that a probe must be present in more than 50% of samples of all 4 phenotypes in order to be kept for further analysis. After this step, 9738 probes were kept for further analysis. Our data selection, collection, rigorous preprocessing and probe filtering produced a potentially valuable data set that can be used by other researchers to identify transcriptomic signatures of various liver diseases.

2.2. Overview of the data set

Figure 1, a-c show the principal component plots of the data set after consensus preprocessing and probe quality filtering. Four different colors represent different phenotypes, and seven different marker shapes represent 7 different experiment batches (batch 1 and batch 2 belong to the same lab). As shown in Figure 1a and Figure 1b, there are two main clusters along the first principal component: batch 7 and batch 1-6. Samples in batch 7 (GSE17183) come from a Japanese population, while the rest batches mostly consist of Caucasian samples. The separation along the first component is likely due to population differences. This poses a serious challenge for chronic hepatitis C, where two populations of comparable size exist: 10 Caucasian samples (27.8%) of and 26 Japanese samples (72.2%). In Figure 1c, where the second principal component is plotted against the third principal component, there is preliminary separation by phenotypes (denoted by four different colors). Our classification method was able to pick up

phenotypic signals despite strong non-phenotypic background and achieved high classification accuracy even for chronic hepatitis C.

2.3. Overview of a coarse-to-fine search algorithm for liver disease classification

We used a method that integrates relative expression reversal with a coarse-to-fine search strategy to do multi-disease classification [11]. This method includes three steps (Figure 2). First, for all possible binary phenotype combinations, the following score is calculated for each possible gene pair combination:

$$\Delta_{i,j} = \left| P(X_i > X_j | \text{class } 1) - P(X_i > X_j | \text{class } 2) \right|$$

Where $P(X_i > X_j | \text{class } 1)$ is the frequency of gene *i* expressed higher than gene *j* in class 1 [8]. The gene pair that achieves the maximum score is used as the top scoring pair (TSP) for this binary phenotype classification. The diagnostic tree construction process is essentially the same as hierarchical clustering, where most similar phenotypes are iteratively combined; instead of distance metric, our method uses the score of the top scoring pair as the metric for phenotype similarity. Second, it finds gene pair classifiers at each node and branch of the diagnostic tree. Classifiers at each node are used to separate diseases at that particular node from all other diseases considered (one vs. the rest); classifiers at each branch are used to separate diseases of the left child node from diseases of the right child node (one vs. one) if classification ties occur. In classification, a sample of unknown class starts from the root node, and moves through classification at each *node* until it reaches a leaf node(s). If it is accepted at multiple nodes, the branch classifier is used to break the tie, since there is only one unique path to a particular leaf starting from the root in this diagnostic tree. The benefit of using the node classifier is that a sample can belong to none of the liver diseases considered and be rejected at any place in the diagnostic process, while using only the branch classifier (a pure decision tree) would force a sample to belong to a unique disease in the tree, which may not be desirable.

2.4. Overview of method to identify deregulated networks in liver pathogenesis

The Differential Rank Conservation (DIRAC) method [15] assesses combinatorial gene interactions to quantify various biological pathways or networks in a comparative sense. This approach is based on the relative expression values of participating genes—i.e., the ordering of expression within network profiles [15]. Using this method, we identified the most variably expressed networks, representing statistically robust differences between disease states at the network level, between all liver diseases and normal liver, as well as between HCC and non-malignant phenotypes. Figure 3 describes the overall steps of the DIRAC method.

2.5. Top-Scoring-Pair based decision tree performs highly accurate classification

To estimate the classification error of this method, we did 10 runs of standard 10-fold cross validation, so we can get both the classification accuracy and variance. The probe filtering procedure, where we eliminated probes that were absent in more than 50% of samples of any phenotype, was done inside

each cross validation loop on the training set only, in order to avoid possible cross talk between training and testing samples. Table 2 shows the confusion matrix of classification results. The diagonal elements represent correct classifications; the off-diagonal elements represent misclassifications. The average accuracy across all classes is 85.2±1.6%, with lowest 72.6±5% (normal liver) and highest 91.7±4.6% (chronic hepatitis C). Class-specific accuracy and accuracy variance is shown in Figure 4. It is important to notice that the classification accuracy for chronic hepatitis C is much higher than what is possible based purely on population difference (27.8% Caucasian vs. 72.2% Japanese), which would only be 72.2% at most. To compare this method with other state of the art classification algorithms, we applied Support Vector Machines (SVM) on the same dataset. We used the libsvm package [26] for SVM. To ensure a fair comparison, the number of features provided in the training set was the same (around 9700, varies in cross validation), and the number of features selected by each method in the training set was also the same (<=40). The classification results are comparable: SVM achieved average accuracy of 87.7±1.9% in 10 runs of 10 fold cross validation. The advantage of our method is that it is a parameter-parsimonious method that uses very few genes to perform highly accurate multi-class classification, while keeps the decision rules straightforward to interpret.

2.6. Classification performance in independent validation

Figure 5 shows the result of leave-one-lab-out independent validations. Our data set consists of 7 different batches from 6 different labs. In each round of leave-one-lab-out validation, we left all samples belonging to a lab for testing while trained the classification method on all the other labs. We processed training and testing samples separately starting from individual raw intensity files: do GCRMA, probe filtering and feature selection only on training labs to ensure truly independent validation. From Figure 5, we can observe that when there are reasonably large numbers of samples (~100 samples) from at least two different labs, we can get comparable accuracy in independent validation as in cross validation (scenario 1). For example, when we when we left out GSE6764, the accuracy for classification remained high for both cirrhosis (131 samples in training from 3 labs) and HCC (138 samples in training from 2 labs), indicating that most of the variance of both phenotypes can be captured by samples in the training set. In another scenario where there are too few samples in training, even there are multiple labs, classification accuracy is much worse (scenario 2). This is the case for hepatocellular carcinoma when we left out GSE9843 (82 samples in training from 2 labs). We almost completely failed when we trained on samples from one population and tested on another (scenario 3). This is the case for cirrhosis and chronic hepatitis C when we trained on Caucasian samples and tested on samples from a Japanese population (GSE17183) or vice versa.

The classifiers presented in our study are based on training on the whole data set. Classification for cirrhosis and hepatocellular carcinoma resembles scenario 1, while classification for normal liver and chronic hepatitis C resembles scenario 2. Based on observations from independent validations, we have good confidence in classifiers related to cirrhosis and hepatocellular carcinoma, and modest confidence

in classifiers related to normal and chronic hepatitis C. As microarrays become widely used in hepatology research, we anticipate that there will be more samples for each major liver pathophysiological state, and disease classifiers produced by our method would be more and more accurate in independent validations given greater availability of these high throughput data.

2.7. Classifier genes are closely involved in liver disease pathogenesis

Through literature search, we found that many genes in our classifier are involved in important biological processes (immune response and inflammation, fibrogenesis, metabolism, and cellular signaling) of liver disease pathogenesis and their up or down regulation in a particular disease may have a mechanistic explanation. Classifier genes information is shown in Table 3. Genes with literature evidence of potential involvement in liver disease are listed in Table 4.

HLA-B, FAM21, EDEM1, UBD and CD74 reflect the presence of hepatitis C infection. Higher expression of HLA-B and CD74 in patients infected with HCV may reflect hosts' immune response to persistent viral infection. CD74's ligand, MIF (macrophage migration inhibitory factor) has a wide range of functions that links inflammation to carcinogenesis: it suppresses immunosurveillance [27], contributes to neoangiogenesis and epithelial cell proliferation [28], and suppresses p53 function [29]. Serum level of MIF has been reported to be elevated in HCC and liver cirrhosis patients [30]. Its overexpression is also associated with poor prognosis in HCC patients [31]. Our meta-analysis showed that higher expression of MIF's cognate receptor CD74 may be a possible mechanism that this cytokine exerts its effects in liver disease pathogenesis. EDEM1 is directly involved in endoplasmic reticulum-associated degradation (ERAD), and its transcription is inhibited by HCV non-structural protein 4B (NS4B) [32], which may explain its lower expression in liver disease patients. UBD is negatively regulated by p53 [33], and it has been reported that HCV core protein relieves p53-mediated suppression [34], which may explain UBD's higher expression in liver diseases of HCV etiology.

AGRN, COL4A1, and ADAMTSL3 may participate in the fibrogenesis process characteristic of liver diseases. As a result of bile ductules proliferation and new blood vessels formation in liver cirrhosis as well as neoangiogenesis in HCC, there is a drastic increase in the quantity of AGRN in HCC and liver cirrhosis [35]. Type IV collagen can stimulate resident hepatic stellate cells, a major collagen producer in liver fibrosis, by activating latent cytokines such as TGF-beta1 [36]. Both plasma and hepatic level of type IV collagen peptide have been reported to be elevated in chronic hepatitis C and cirrhosis patients [37-39]. ADAMTSL3 is a member of the ADAMTS family of proteins. ADAMTS proteins are known to be involved in collagen processing, cleavage of the matrix proteoglycans, and anti-angiogenesis [40]. ADAMTSL3 is located in the extracellular matrix, and may be involved in cell-matrix interactions or assembly of specific extracellular matrices [41].

ACAT1, DUT, MT1G, GLRX5, and AGPAT1 are involved in different aspects of metabolism that may be related to liver disease pathogenesis. As HCV infection results in increased fatty acid synthesis [42] and inhibition of fatty acid oxidation [43], the reduced expression of ACAT1 in liver diseases of HCV etiology

may reflect hosts' response to decreased demand of ketone body metabolism. DUT encodes an essential enzyme of nucleotide metabolism that hydrolyzes dUTP to dUMP, a precursor for the synthesis of thymine nucleotides needed for DNA synthesis. DUT upregulation may be the result of hepatocyte regeneration in chronic hepatitis C and cirrhosis, and tumor cell proliferation in HCC, both of which require DUT activity for increased DNA synthesis relative to resting hepatocyte in healthy liver. MT1G binds to heavy metals and reduces cellular oxidative stress by capturing harmful oxidant radicals like the superoxide and hydroxyl radicals [44]. A significant inverse correlation between MT protein or mRNA expression and both the grade and the stage of fibrosis has been reported [45]. Its expression is significantly decreased in HCC and hepatoblastoma as a result of promoter hypermethylation [46, 47]. GLRX5 is a member of the glutaredoxins that maintain the cellular redox state, serving as essential protein antioxidants [48]. GLRX5 deficiency in human is characterized by anemia and iron overload [49]. In the liver, excess iron acts as a pro-inflammatory agent and is associated with increased morbidity and mortality of HCV related chronic liver disease [50]. Down-regulation of these two redox-balancing proteins may thus exacerbate the increased cellular oxidative stress caused by HCV [51] and contribute to disease progression. AGPAT1 converts lysophosphatidic acid (LPA) to phosphatidic acid (PA) [52]. The bioactive LPA is a potent mediator of tissue repair and wound healing [53, 54], and wound healing is a characteristic process in liver fibrosis and cirrhosis as a result of chronic liver injury afflicted by HCV [55]. While LPA increased DNA synthesis and MAP kinase activity in hepatic stellate cells, the major player in liver fibrogenesis, it decreased DNA synthesis in hepatocytes [56]. AGPAT1's lower expression in cirrhosis relative to HCC may reflect LPA's decreased conversion to PA and increased involvement in the fibrogenesis process; while its higher expression may reduce LPA level in HCC, and LPA has a negative effect on hepatocyte proliferation.

TACSTD2, MAPKAPK2, ADIPOR1, ALPL, ENDRB, and ANXA3 may participate in various cellular signaling events that may contribute to liver disease pathogenesis. TACSTD2's higher expression in cirrhosis relative to HCC may reflect the oval cell activation event in liver cirrhosis [57]. MAPKAPK2's higher expression in HCC relative to cirrhosis may contribute to HCC invasion. ADIPOR1 encodes a receptor for adiponectin, a protein hormone that modulates a number of metabolic processes, including glucose regulation and fatty acid catabolism [58]. Interestingly, in a 9-year follow-up study, chronic hepatitis C and cirrhosis patients with higher serum adiponectin had a higher incidence of developing HCC [59]. EDNRB encodes a non-specific receptor for endothelin 1, 2, and 3. Its relative expression in cirrhosis and HCC is consistent with our meta-analysis result. ANXA3 have been identified as a potential angiogenic factor that induces VEGF production through HIF-1 pathway [60], and is induced during hepatocyte regeneration, a characteristic process in cirrhotic nodules [61].

From the above discussion, we can see that many genes in our classifier have clear concrete links to liver disease pathogenesis and our method identified these genes as accurate disease classifiers without any prior knowledge.

2.8. Perturbed pathways in liver diseases

Using differential rank conservation method, we did two comparisons: all liver diseases vs. normal live (Table 5), and HCC vs. non-malignant liver phenotypes (Table 6). The first comparison is to identify commonly perturbed pathways in liver disease pathogenesis, and the second comparison is to identify perturbed pathways in hepatocellular carcinogenesis.

The commonly perturbed pathways in liver diseases identified in our analysis have been previously implicated in liver disease pathogenesis. The CXCL12/CXCR4 pathway is implicated in recruiting liver infiltrating lymphocytes [62] and is the key player in HCV-associated liver inflammation and fibrosis [63]. EGF up-regulation is a characteristic event in both cirrhotic liver disease and HCC [64, 65], and there is mounting evidence supporting its role in linking inflammation caused by chronic liver injury like HCV infection, liver fibrosis, and cancer [66]. The mTOR pathway plays a pivotal role in HCC [67] and links inflammation to tumor angiogenesis [68]. Aberrant mTOR signaling in liver is associated with both IGF1 and EGF pathway activation [67], both of which are also listed among the top 10 perturbed pathways. Both the HIVNEF and the TNFR1 pathways may be related to the ability of HCV core protein to inhibit apoptosis mediated by Fas-L and tumor necrosis factor alpha [69, 70]. HCV is known to impair insulin signaling and induce insulin resistance by down-regulating two key components in this pathway: IRS1 and IRS2 [71], and provides a mechanistic explanation for increased prevalence of type 2 diabetes in chronic hepatitis C patients [72]. Both the PLATELETAPP pathway and the INTRINSIC pathway are both involved in blood coagulation. The liver plays a central role in the blood coagulation process as the major synthesis site for coagulation factors, and acute and chronic liver diseases are invariably associated with coagulation disorders [73].

Among the top 10 variably expressed pathways between HCC and non-malignant phenotypes, the integrin and the extracellular matrix (ECM) signaling pathways are ranked 1st and 10th respectiviely, pointing out the crucial role of tumor microenvironment in hepatocellular carcinogenesis. The fibrotic microenvironment in the liver is characterized by an altered composition of the extracellular matrix (ECM), and key components of the integrin and ECM signaling are crucial regulators of HCC invasion [74]. The MAPK pathway is activated by multiple growth factors and its increased activation in HCC has already been identified [75]. The serum level of IL-6, a pro-inflammatory cytokine, is significantly higher in HCC compared to cirrhosis and control [76]. Both liver injury and compensatory proliferation were strongly dependent on IL-6 and that the absence of this tumor promoting cytokine resulted in almost complete inhibition of DEN-induced hepatocarcinogenesis [77]. According to Biocarta pathway definition, the HER2 partway consists of components from the IL-6 signaling pathway, the ERK pathway, and the AKT pathway, all of which play crucial roles in hepatocellular carcinogenesis. Therefore, the HER2 pathway may reflect the synergistic effects of these three pathways. The c-MET pathway mediates the signaling of hepatocyte growth factor (HGF), a potent mitogen for hepatocytes. This c-MET is overexpressed in HCC, and higher expression is associated with poor 5-year survival [78].

It is interesting to notice that many classifier genes and the variably expressed pathways point to similar biological processes in liver disease pathogenesis. HLA-B and CD74 are involved in immune response to hepatitis C virus and ensuing inflammation and carcinogenesis; the CXCR4 and IL-6 pathways also participate in this process. AGRN, COL4A1, and ADMATSL3 are either basement membrane components or mediate cell-matrix interaction important in liver fibrosis; the integrin and ECM signaling pathways also participate in this process. COL4A1 is also part of the PLATELETAPP pathway and the INTRINSIC pathway, both of which are commonly perturbed in liver diseases. ACAT1, DUT, and AGPAT1 are involved in fatty acid, nucleotide, and phospholipids metabolism respectively; signaling through the IGF-mTOR pathway results in decreased fatty acid oxidation [79], increased lipid synthesis [80], and DNA synthesis. MAPKAPK2 is a downstream target of the MAPK pathway, which is ranked as the 4th variably expressed pathway between HCC and non-malignant liver phenotypes.

2.9. Tables

Table 1. Dataset and phenotype sample distribution

		Phenotypes				sum
		Normal	CHC	Cirrhosis	HCC	
GEO accession	GSE14323	19	0	58	47	124
	GSE17967	0	0	63	0	63
	GSE6764	10	0	13	35	58
	GSE9843	0	0	0	91	91
	GSE11190	2	10	6	0	18
	GSE14668	8	0	0	0	8
	GSE17183	0	26	4	0	30
sum		39	36	144	173	392

CHC: chronic hepatitis C; HCC: hepatocellular carcinoma

Table 2. Confusion matrix of classification performance

		Predicted disease classes			
		Normal	**CH**	**HCC**	**CHC**
Actual disease classes	**Normal**	**72.6%**	3.8%	18.5%	5.1%
	CH	0.2%	**91.0%**	4.9%	4.0%
	HCC	0.1%	14.0%	**85.5%**	0.5%
	HCV	3.1%	2.5%	2.8%	**91.7%**

CH: cirrhosis; HCC: hepatocellular carcinoma; CHC: chronic hepatitis C

Table 3. Classifier genes information

a. List of node-based classifier genes

Disease	gene i		gene j	
	gene symbol	probe	gene symbol	probe
CH,HCC,CHC	HLA-B	209140_x_at	ACAT1	205412_at
	FAM21	212370_x_at	EDEM1	203279_at
	PBXIP1	214177_s_at	ADAMTSL3	213974_at
	UBD	205890_s_at	SLC17A2	207097_s_at
	AGRN	212285_s_at	ALPL	215783_s_at
	DUT	209932_s_at	MFAP3L	205442_at
	CD74	209619_at	MT1G	204745_x_at
CH,HCC	COL4A1	211980_at	ASH2L	209517_s_at
	STX16	221499_s_at	TMEM57	218562_s_at
	CAP1	200625_s_at	GLRX5	221932_s_at
CHC	FAM149A	214890_s_at	DHX40	218277_s_at
CH	CLDN10	205328_at	AGPAT1	32836_at
	TACSTD2	202286_s_at	HGS	210428_s_at
	EDNRB	206701_x_at	PPP2R5D	202513_s_at
	GPM6A	209469_at	FLAD1	205661_s_at
	ANXA3	209369_at	MAPKAPK2	201461_s_at
	SH3YL1	204019_s_at	ADIPOR1	217748_at
	MYO10	201976_s_at	MRPL2	218887_at
HCC	AGPAT1	32836_at	CLDN10	205328_at
	HGS	210428_s_at	TACSTD2	202286_s_at
	PPP2R5D	202513_s_at	EDNRB	206701_x_at
	FLAD1	205661_s_at	GPM6A	209469_at
	MAPKAPK2	201461_s_at	ANXA3	209369_at
	ADIPOR1	217748_at	SH3YL1	204019_s_at
	MRPL2	218887_at	MYO10	201976_s_at

CH: cirrhosis; HCC: hepatocellular carcinoma; CHC: chronic hepatitis C

b. Branch-based classifier information

Disease separation	gene i	probe i	gene j	probe j
CH,HCC vs. CHC	DHX40	218277_s_at	FAM149A	214890_s_at
CH vs. HCC	CLDN10	205328_at	AGPAT1	32836_at

CH: cirrhosis; HCC: hepatocellular carcinoma; CHC: chronic hepatitis C

Table 4. Genes with literature evidence of potential involvement in liver disease

Gene symbol	Involvement in liver disease	Reference
HLA-B	antigen presentation; higher expression in HCC	[81]
CD74	antigen presentation; receptor for macrophage migration inhibitory factor	[82, 83]
EDEM1	targets misfolded glycoprotein for degradation; inhibited by HCV NS4B	[32, 84]
UBD	overexpressed in HCC; contributes to genome instability	[85, 86]
AGRN	liver basement membrane component, accumulated in cirrhosis and HCC	[35]
COL4A1	liver basement membrane component; elevated level associated with liver fibrosis	[37-39, 55]
ADAMTSL3	highly expressed in liver; extracellular matrix component	[41]
ACAT1	catalyzes formation of acetoacetyl-CoA, key step in ketone metabolism	[87]
DUT1	overexpressed in proliferating hepatomas; associated with poor survival in HCC	[88, 89]
MT1G	reduces oxidative stress; down-regulated in HCV-infected patients and HCC	[44-46]
GLRX5	redox balance; deficiency results in iron overload, a risk factor for liver disease	[48-50]
AGPAT1	converts lysophosphatidic acid, a stimulator of wound healing, to phosphatidic acid	[52, 53]
TACSTD2	marker for hepatic oval cell, a facultative hepatic stem cell type	[90]
MAPKAPK2	activated by MAPK, involved in cancer cell invasion	[91-93]
ADIPOR1	receptor for adiponectin, a risk factor for developing HCC	[59, 94, 95]
ALPL	increased serum level reflects malignant infiltration in liver	[96]
ENDRB	overexpressed in cirrhosis; reduced in HCC due to promoter hypermethylation	[97-99]
ANXA3	potential angiogenic mediator; increased expression in liver regeneration	[60, 61]

Table 5. Variably expressed networks between liver diseases and normal liver

Network name	Num.genes	Num.gene pairs	Apparent accuracy	*p*-value
CXCR4_PATHWAY	17	136	0.915	<7E-07
EGF_PATHWAY	21	210	0.908	<7E-07
HIVNEF_PATHWAY	41	820	0.9	<7E-07
IGF1MTOR_PATHWAY	17	136	0.879	<7E-07
INTEGRIN_PATHWAY	25	300	0.878	<7E-07
IGF1_PATHWAY	16	120	0.874	<7E-07
TNFR1_PATHWAY	20	190	0.874	<7E-07
PLATELETAPP_PATHWAY	9	36	0.868	<7E-07
INTRINSIC_PATHWAY	19	171	0.867	<7E-07
INSULIN_PATHWAY	16	120	0.865	<7E-07

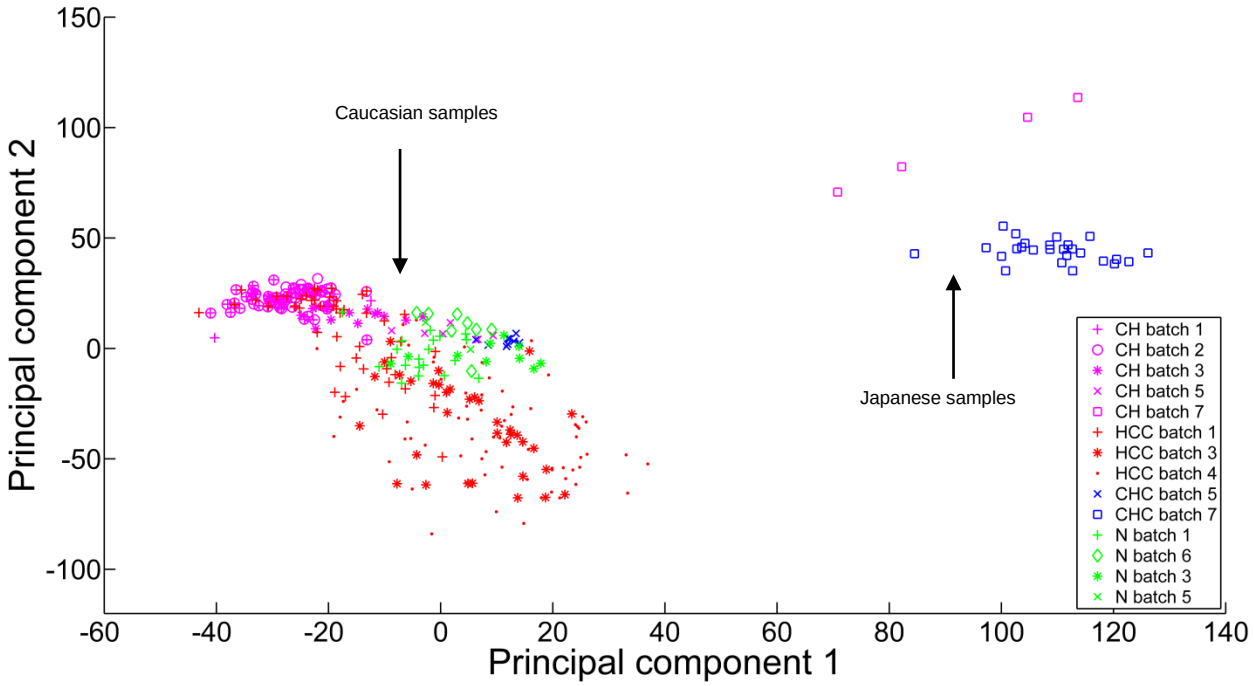
Table 6. Variably expressed networks between HCC and non-malignant phenotypes

Network name	Num.genes	Num.gene pairs	Apparent accuracy	*p*-value
INTEGRIN_PATHWAY	25	300	0.84	<7E-07
HER2_PATHWAY	17	136	0.83	<7E-07
EDG1_PATHWAY	17	136	0.825	<7E-07
MAPK_PATHWAY	54	1431	0.823	<7E-07
IGF1R_PATHWAY	18	153	0.821	<7E-07
MET_PATHWAY	22	231	0.818	<7E-07
IL6_PATHWAY	17	136	0.817	<7E-07
CXCR4_PATHWAY	17	136	0.815	<7E-07
AT1R_PATHWAY	23	253	0.814	<7E-07
ECM_PATHWAY	18	153	0.813	<7E-07

2.10. Figures

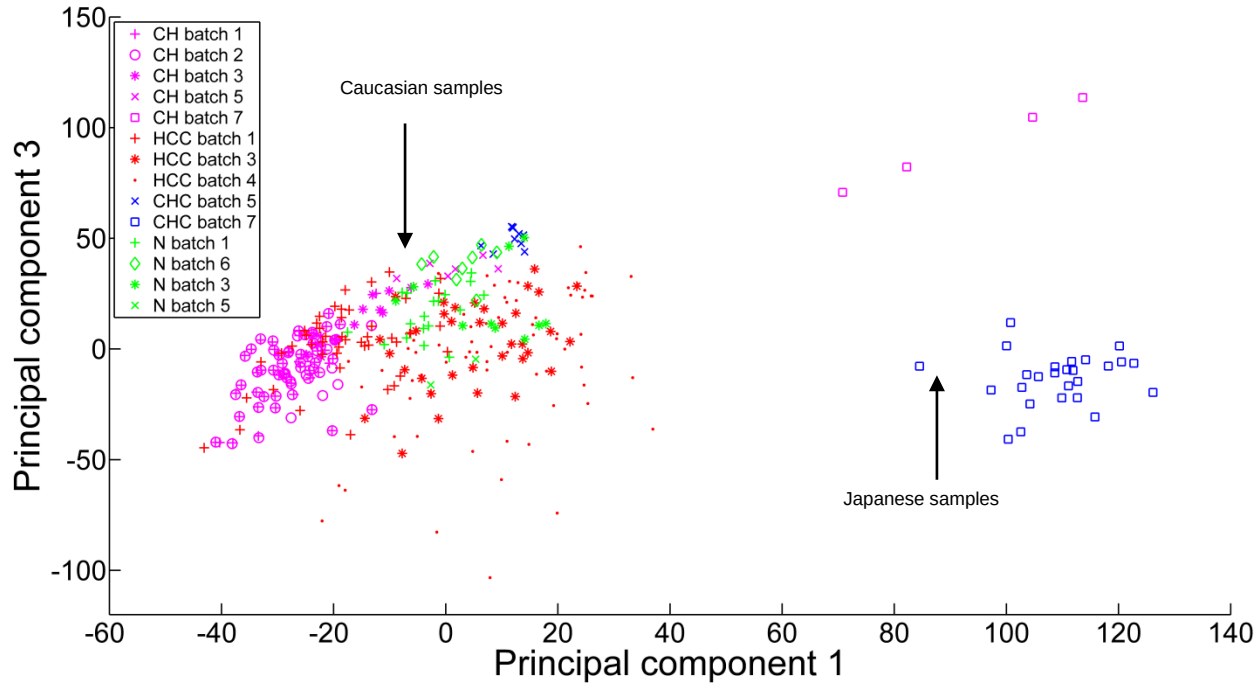
Figure 1. Principal component plots of processed dataset

1.a. 1st vs. 2nd



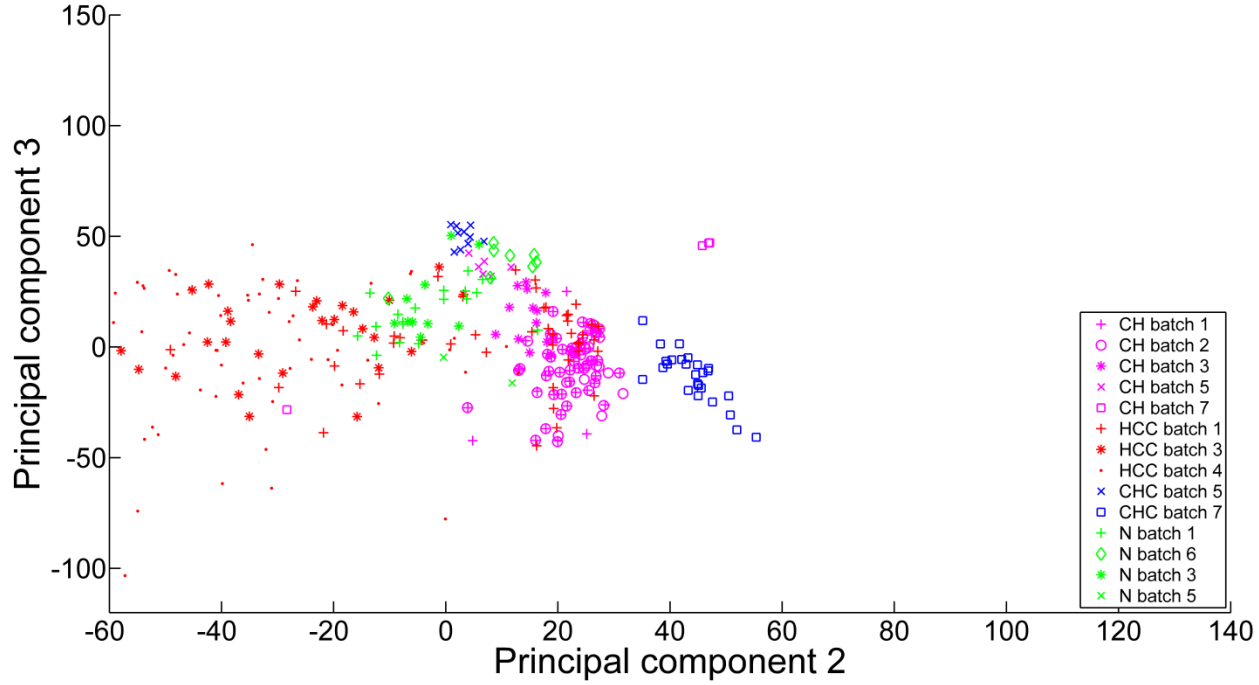
CH: cirrhosis; HCC: hepatocellular carcinoma; CHC: chronic hepatitis C; N: normal liver

1.b. 1st vs. 3rd



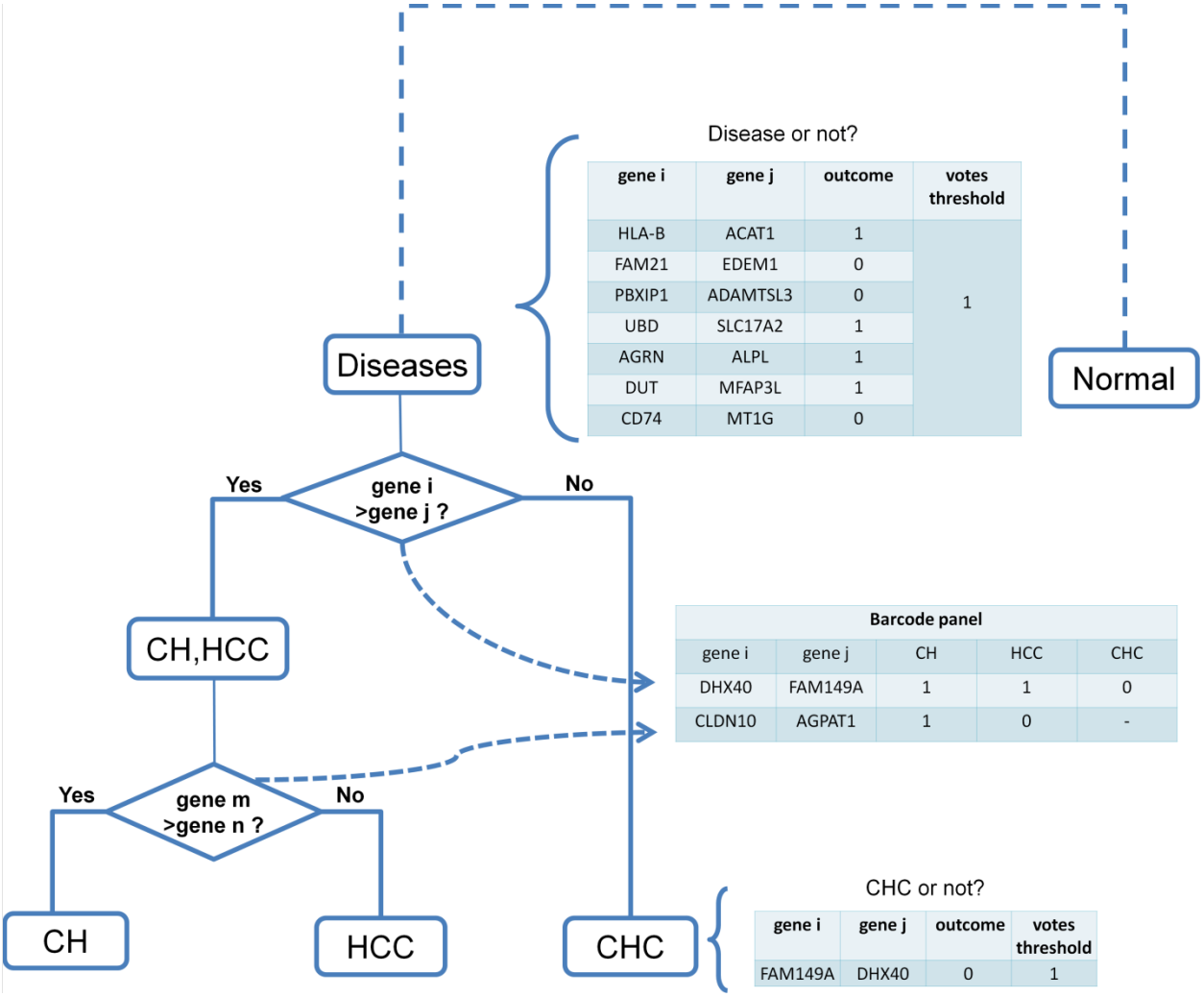
CH: cirrhosis; HCC: hepatocellular carcinoma; CHC: chronic hepatitis C; N: normal liver

1.c. 2nd vs. 3rd



CH: cirrhosis; HCC: hepatocellular carcinoma; CHC: chronic hepatitis C; N: normal liver

Figure 2. Diagnostic tree construction, deification of node and branch based classifiers, and classification of sample



The diagnostic tree is constructed through an essentially hierarchical clustering approach where the score of the top scoring pair is used as a similarity metric. The top scoring pair at each branch and the binary outcome of each phenotype at each branch is accumulated into a barcode panel, where each phenotype has a unique binary signature. The barcode panel is used to resolve ties in classification. Classifiers at each node are used to decide if a sample belongs to the phenotypes at that node based on an optimal threshold. Normal liver is the default classification class. It is not used in tree construction, but used as a negative set to train classifiers for the "Diseases" node. As illustrated in the figure, the hypothetical sample would be accepted at the "Diseases" node since it passed the vote threshold, and proceeds to *both* the "CH, HCC" node and the "CHC" node. It would be rejected at the "CHC" node since it did not pass the vote threshold at that node.

Figure 3. Overview of the Differential Rank Conservation (DIRAC) methods. Adopted with permission from Eddy *et al* [15].

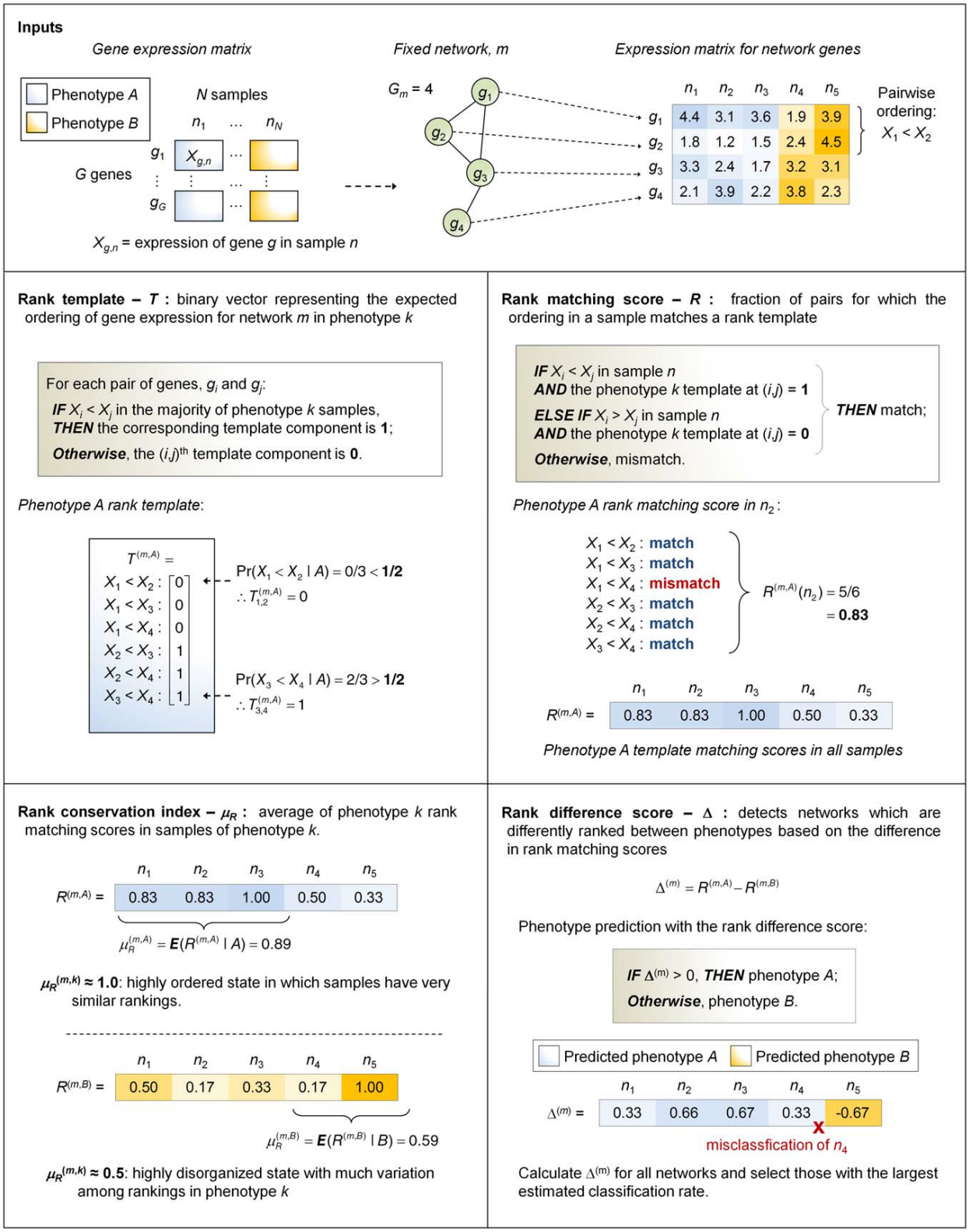
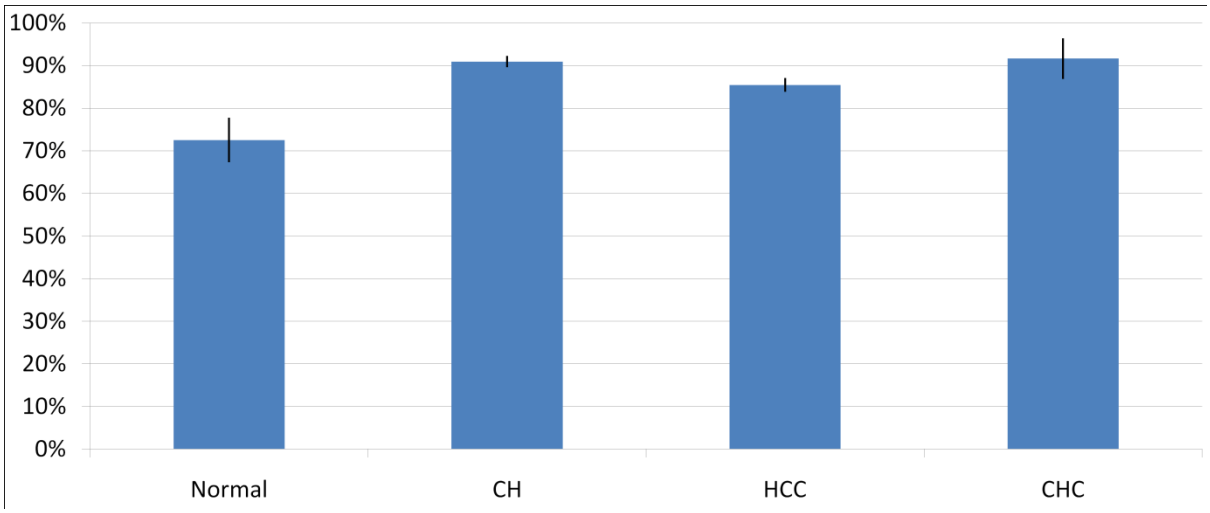
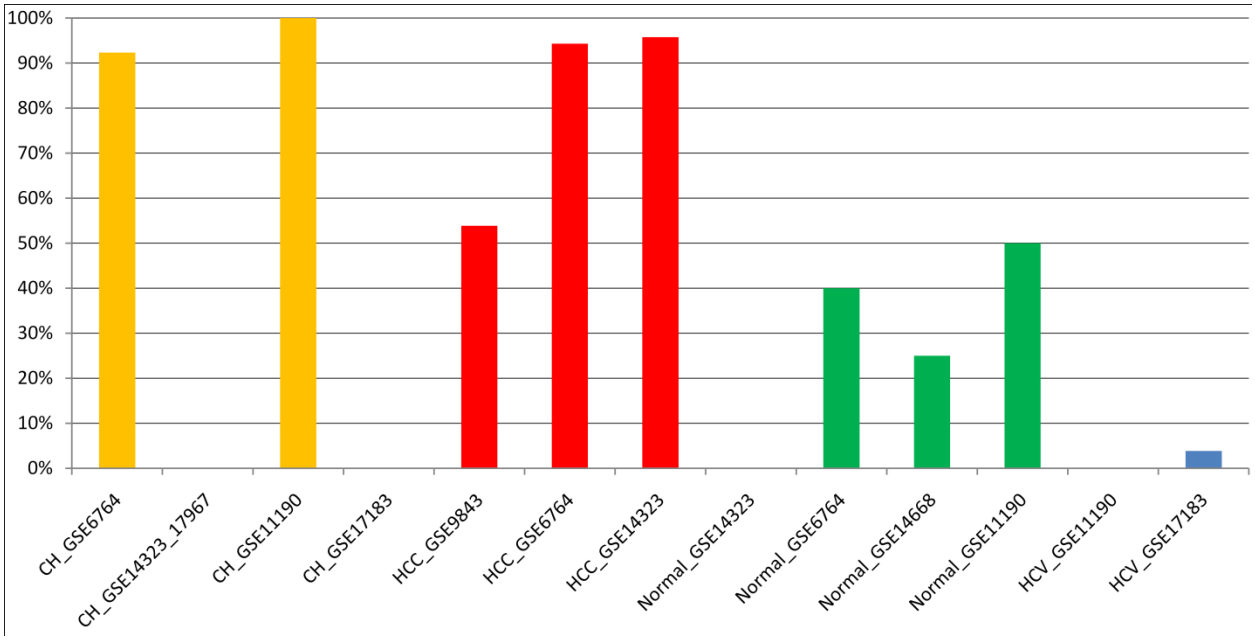


Figure 4. Disease classification accuracy and variance in 10 runs of 10 fold cross validation



CH: cirrhosis; HCC: hepatocellular carcinoma; CHC: chronic hepatitis C

Figure 5. Leave-one-lab out independent validation results



CH: cirrhosis; HCC: hepatocellular carcinoma; CHC: chronic hepatitis C

CHAPTER 3

CONCLUSION

In this study, we collected, processed and analyzed 392 microarray samples of 4 liver pathophysiological states from 7 published studies. We took a comprehensive approach to consider all 4 major liver pathophysiological states in a single unified context, and identified specific transcriptomic signatures for each phenotype. Based on gene pairs whose relative orderings are most informative of phenotype distinction, we built a coarse-to-fine diagnostic tree from bottom up, which performed classification at each node and leaf of the tree. Our classifier consists of 36 unique features, and has 85.2±1.6% average accuracy across 10 runs of 10 fold cross-validation. Our classifier has comparable accuracy with SVM (87.7±1.9%), when both methods used the same number of features for training, and selected similar number of features in their respective classifiers. The main advantage of our method is that it is a data-driven, parameter-parsimonious method that requires little tuning from the user. Classification is based on relative ranking of gene expression values, not absolute intensities, thus this method is invariant to normalizations that preserve relative ordering of genes. Another benefit is that it produces more interpretable classifiers (a list of "gene *i* expressed higher than gene *j* " in each disease) than SVM, and may be more conducive to production of diagnostic tests.

We did stringent leave-one-lab out independent validations to validate our method, and found that as long there are samples from two independent labs with reasonable total sample size (around 100 samples) to train our classification method, highly accurate classification can be achieved on a completely independent test data set without any data integration. This is probably because accurate classifiers trained from two independent labs captured the true characteristics of the phenotype, instead of any particular lab. There are many data sets on Gene Expression Omnibus that studied the same disease on same or similar platforms (e.g. Affymetrix U133A, U133 2.0 or U133 Plus 2.0), and more data sets will be available with the wide-spread use of DNA microarray and other high throughput technology like RNA-seq. The results of our independent validations suggest that properly integrating samples from multiple sources may be a potentially useful way to mitigate the lab or batch effects inherent in these high throughput data.

We found that many genes in our classifier are involved in important biological processes of liver disease pathogenesis. They participate in immune response, inflammation, oxidative stress response, fibrogenesis, and metabolism. Many genes' relatively higher or lower expression in a particular phenotype can be explained by mechanisms such as direct inhibition by HCV proteins, indirect activation by HCV proteins, and promoter hypermethylation. Using a new method developed in our lab, we also identified commonly perturbed networks in liver diseases compared to normal liver, and perturbed networks in HCC compared to non-malignant phenotypes. We were able to identify networks known to play important roles

in liver disease pathogenesis: CXCR4, mTOR, epidermal growth factor, insulin growth factor 1, and apoptosis-related signaling networks. These pathways link immunity, inflammation, fibrosis, metabolism and malignant transformation. Perturbed networks in HCC reflected two important players in hepatocarcinogenesis: dysregulation of growth factor signaling (EGF, MAPK, IGF, and MET) [100] and tumor microenvironment (integrin and ECM signaling) [101]. Both the classifiers and the perturbed pathways identified in this study offer biological insight into the pathogenesis of liver diseases, and may be potential candidates for novel diagnostic approaches to multiple liver diseases.

CHAPTER 4

METHODS

4.1. Building a coarse-to-fine diagnostic tree and identifying branch-based classifiers

Let E be a P×N matrix of P genes and N samples. Let C= {C_1, C2,...C_M} be the set of possible class labels. In this study, there are 4 phenotypes, so M=4. C_4 represent normal liver, the default class a sample would be classified, and would not be in the iterative tree-building process. First, we rank the gene expression values within each sample. After this step, the absolute expression intensities are replaced by relative orderings, the expression matrix E becomes a rank matrix X. Let $P_{ij}(C_m) = Pr(X_i > X_j \mid Y = C_m)$ be the possibility of observing $X_i > X_j$ (the expression level of gene i is higher than that of gene j) in class C_m. $P_{ij}(C_m)$ is estimated by the frequency of observing $X_i > X_j$ in class C_m. Let $\Delta_{ij} = |P_{ij}(C_1) - P_{ij}(C_2)|$ be the score of each gene pair (i, j), which quantifies the difference in probability of observing $X_i > X_j$ between class 1 and class 2. A score of 0 means that the ordering $X_i > X_j$ is equally likely in both classes, and the relative ordering of gene pair (i, j) is not informative of class distinction; a score of 1 means we always observe $X_i > X_j$ in class 1 and never in class 2, and the relative ordering of gene pair (i, j) is highly informative of class distinction. The higher the score, the better gene pair (i, j) can classify class 1 and 2. The score of the gene pair that can most accurately classify class 1 and 2 is denoted as Δ_{max}, and the gene pair is called "top scoring pair" (TSP) [8]. To reduce our search space and avoid potential over-fitting, we selected the top 2500 differentially expressed genes with Wilcoxon rank sum test inside cross validation for TSP calculation. The calculation of TSP from around thousands of genes, representing millions of possible binary combinations is a computationally intensive yet parallelizable task. Therefore, we used the GPU-implementation of TSP calculation which resulted in dramatic speedup [102].

Building the coarse-to-fine diagnostic tree is the same as hierarchical clustering, where the most similar phenotypes are iteratively combined to form a new node, and the score of the top scoring pair is used as the similarity metric. The top scoring pairs at each bifurcation of the diagnostic tree are accumulated into a barcode panel, where each disease class has a unique binary signature recording the outcome of whether $X_i > X_j$ at each bifurcation of the tree.

4.2. Identifying node-based classifiers

The major difference our diagnostic tree and a pure decision tree is that classification happens at the nodes of the tree, not at the branches (bifurcations). In a pure decision tree, once a sample is inside the tree, it is forced to go to a unique leaf. However, in our diagnostic tree, starting from the root, a sample will go to *both* child nodes of the current node, and can be accepted at neither of the child nodes, either of the child nodes, or both of the child nodes. The branch-classifiers are only used when there are multiple diagnosis: a sample is accepted at multiple leaves of the diagnostic tree. The advantage of this feature is

that even a sample is inside our diagnostic tree, it can still be rejected at any node of the tree, and does not have to be diagnosed as one of the classes inside the tree. The basis of this important feature is the classifiers at each node of the diagnostic tree.

When identifying the node-based classifiers, we already have the structure of the diagnostic tree. At each node, there is a positive and a negative set: positive set is the set of classes belonging to that node, negative set is the set of classes whose members reach that node yet do not belong to the positive set. This is possible because samples may be misclassified at the previous nodes and make to both of the child nodes. At the root node, the positive set is all 3 liver diseases, and the negative set is normal liver. Let x = (1, 3, 5, 7, 9) be the maximum possible number of top scoring gene pairs (not only the best gene pairs, but the 2nd best, 3rd best, and so on) to be used classify the positive set and negative set. Let K be the number of top scoring pairs actually used. Let k be the number of "votes" a sample needs in order to be accepted at that node ($k \leq K \leq x$). An internal cross-validation loop is performed to identify the optimal K and k for each node that maximize classification sensitivity.

4.3. Classification of microarray samples

Classification starts at the root node, and proceeds to both of the child nodes. Let μ denote the number of votes (how many times $X_i > X_j$ is true for that sample for all K top scoring pairs used at that node) a sample gets. If $\mu < k$ (the threshold), then the sample would be rejected, and classified as the default class: normal liver. If $\mu \geq k$, the sample would be accepted at the current node, and proceeds to both of its child nodes. The previous process is repeated until the sample reaches a leaf (leaves). If the sample is accepted at multiple leaves, the branch-based classifiers, which essentially form a pure decision tree, are used to break the tie. Since there is only one unique path from the root to any leaf, a unique single diagnosis is guaranteed.

4.4. Identifying perturbed networks

All steps described in this part is based on the original DIRAC method [15].

First, we downloaded network definitions from Biocarta (Mar 30, 2011). Given the list $\{g_1, ..., g_{Gm}\}$ of G_m genes within a network m on a microarray, let $\mathbf{X}$ denote the corresponding expression profile, where X_i is the expression profile of gene g_i. Our microarray data then consists of a $G_m \times N$ matrix; the n^{th} column represents the expression profile $\mathbf{x}_n$ of the nth sample, n=1, ..., N. In addition, each sample is labeled by a phenotype $k \in C=\{C_1,C_2,...C_K\}$. In our case, M=4. DIRAC is based entirely on the ranks within each expression profile. For a network of G_m genes, it considers all $G_m (G_m -1)/2$ possible orderings of all gene pairs. For example, if there are =4 genes, then there are six distinct ordered pairs: {(1, 2)}, {(1, 3)}, {(1, 4)}, {(2, 3)}, {(2, 4)}, {(3, 4)}.

Rank template matching for networks. In this step, we define a template representing the expected ranking of network genes within a phenotype. We consider $P(C_k) = Pr(X_i > X_j \mid Y = C_k)$ for each pair of genes (i, j) of network m, and all phenotype C_k. We estimate these probabilities by the observed

frequency of $X_i < X_j$ (the expression level of gene i is higher than that of gene j) in class C_k. The rank template for a fixed network m and phenotype C_k is the binary vector $T^{(m, Ck)}$ of length G_m $(G_m -1)/2$ where the i, j^{th} component is 1 if $P(C_k) >0.5$ and 0 if $P(C_k) <0.5$. Given a sample's expression profile $\mathbf{x}_n$ for the network m, rank matching score $R^{(m, Ck)}(\mathbf{x}_n)$ measures how well the sample matches the template $T^{(m, Ck)}$. $R^{(m, Ck)}(\mathbf{x}_n)$ is defined to be the fraction of the G_m $(G_m -1)/2$ pairs for which the observed ordering within $\mathbf{x}_n$ matches the template: the expected ordering for phenotype C_k.

Rank difference scores. Rank difference is defined as $\Delta^{(m)}(\mathbf{x}_n) = R^{(m, Ck1)}(\mathbf{x}_n) - R^{(m, Ck2)}(\mathbf{x}_n)$, that is, the difference between the rank matching score when sample n is matched to the rank template of phenotype C_{k1} and the rank matching score when sample n is matched to the rank template of phenotype C_{k2}. Clearly, $-1 \leq \Delta^{(m)}(\mathbf{x}_n) \leq 1$ with positive values providing evidence that the phenotype of sample n is C_{k1} and negative values providing evidence that the phenotype of sample n is C_{k2}. Therefore, the difference score provides a classifier for phenotype identification based on the degree of regulation of the genes in network m: phenotype $Y= C_{k1}$ if $\Delta^{(m)}(\mathbf{x}_n) >0$ and phenotype $Y= C_{k2}$ if $\Delta^{(m)}(\mathbf{x}_n) \leq 0$. A good network classifier would have the following property: the variance of $R^{(m, Ck1)}(\mathbf{x}_n)$ of all samples belonging to phenotype C_{k1} and the variance of $R^{(m, Ck2)}(\mathbf{x}_n)$ of all samples belonging to phenotype C_{k2} are both small; the difference between rank templates $T^{(m, Ck1)}$ and $T^{(m, Ck2)}$ is big. In other words, the rank matching scores for this network have low variance within each phenotype, but high variance across phenotypes. Networks with such properties are called "variably expressed networks", and represent statistically robust differences between phenotypes. The accuracy of the network classifier m is defined as the average sensitivity and specificity of identifying relevant phenotypes.

Significance testing for the classification rate of variably expressed networks. After identifying variably expressed networks between two phenotypes based on their ability to accurately classify samples, a permutation test is used to evaluate the significance of these networks. Under the null hypothesis that no systematic difference in gene expression profiles exist between C_{k1} and C_{k2}, (i) the original sample labels were randomly re-assigned to samples, while keeping the number of samples belonging to each phenotype same as the original data set; (ii) sample classes in the permuted data set were predicted as C_{k1} or C_{k2} based on whether the rank difference score was positive or negative, respectively, and scores were assigned to each network as measured by the estimated classification accuracy; (iii) the first two steps were repeated for 10,000 permutations to generate a null distribution of network classification rates; and (iv) the significance level for the classification accuracy of network m was measured as the probability of observing classification rates greater than or equal to the original classification accuracy in the null distribution.

REFERENCES

1. Lavanchy, D., *The global burden of hepatitis C.* Liver International, 2009. **29**: p. 74-81.
2. Farazi, P.A. and R.A. DePinho, *Hepatocellular carcinoma pathogenesis: from genes to environment.* Nat Rev Cancer, 2006. **6**(9): p. 674-87.
3. Schuppan, D. and N.H. Afdhal, *Liver cirrhosis.* The Lancet, 2008. **371**(9615): p. 838-851.
4. Llovet, J.M., A. Burroughs, and J. Bruix, *Hepatocellular carcinoma.* Lancet, 2003. **362**(9399): p. 1907-1917.
5. El-Serag, H.B. and L. Rudolph, *Hepatocellular carcinoma: Epidemiology and molecular carcinogenesis.* Gastroenterology, 2007. **132**(7): p. 2557-2576.
6. Thorgeirsson, S.S. and J.W. Grisham, *Molecular pathogenesis of human hepatocellular carcinoma.* Nat Genet, 2002. **31**(4): p. 339-46.
7. Lemmer, E.R., S.L. Friedman, and J.M. Llovet, *Molecular Diagnosis of Chronic Liver Disease and Hepatocellular Carcinoma: The Potential of Gene Expression Profiling.* Semin Liver Dis, 2006. **26**(04): p. 373-384.
8. Geman, D., et al., *Classifying gene expression profiles from pairwise mRNA comparisons.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article19.
9. Tan, A.C., et al., *Simple decision rules for classifying human cancers from gene expression profiles.* Bioinformatics, 2005. **21**(20): p. 3896-904.
10. Lin, X., et al., *The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations.* BMC Bioinformatics, 2009. **10**: p. 256.
11. Sung, J., *Gene-pair expression signatures accurately reflect health and disease status of human brain.* In preparation
12. Rhodes, D.R. and A.M. Chinnaiyan, *Integrative analysis of the cancer transcriptome.* Nat Genet, 2005. **37 Suppl**: p. S31-7.
13. Warnat, P., R. Eils, and B. Brors, *Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes.* BMC Bioinformatics, 2005. **6**.
14. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nat Rev Genet, 2010. **11**(10): p. 733-739.
15. Eddy, J.A., et al., *Identifying Tightly Regulated and Variably Expressed Networks by Differential Rank Conservation (DIRAC).* PLoS Comput Biol, 2010. **6**(5): p. e1000792.
16. Mas, V.R., et al., *Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma.* Mol Med, 2009. **15**(3-4): p. 85-94.
17. Farci, P., et al., *B cell gene signature with massive intrahepatic production of antibodies to hepatitis B core antigen in hepatitis B virus–associated acute liver failure.* Proceedings of the National Academy of Sciences, 2010. **107**(19): p. 8766-8771.
18. Archer, K.J., et al., *Identifying genes for establishing a multigenic test for hepatocellular carcinoma surveillance in hepatitis C virus-positive cirrhotic patients.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(11): p. 2929-32.
19. Wurmbach, E., et al., *Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma.* Hepatology, 2007. **45**(4): p. 938-47.
20. Chiang, D.Y., et al., *Focal gains of VEGFA and molecular classification of hepatocellular carcinoma.* Cancer Res, 2008. **68**(16): p. 6779-88.
21. Sarasin-Filipowicz, M., et al., *Interferon signaling and treatment outcome in chronic hepatitis C.* Proc Natl Acad Sci U S A, 2008. **105**(19): p. 7034-9.
22. Honda, M., et al., *Differential interferon signaling in liver lobule and portal area cells under treatment for chronic hepatitis C.* J Hepatol, 2010. **53**(5): p. 817-26.

23. Villanueva, A., et al., *Genomics and signaling pathways in hepatocellular carcinoma.* Seminars in Liver Disease, 2007. **27**(1): p. 55-76.
24. Wu, Z.J., et al., *A model-based background adjustment for oligonucleotide expression arrays.* Journal of the American Statistical Association, 2004. **99**(468): p. 909-917.
25. Liu, W.-m., et al., *Analysis of high density expression microarrays with signed-rank call algorithms.* Bioinformatics, 2002. **18**(12): p. 1593-1599.
26. Chang, C.-c. and C.-J. Lin. *LIBSVM: a Library for Support Vector Machines*. 2001; Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm
27. Abe, R., et al., *Regulation of the CTL Response by Macrophage Migration Inhibitory Factor.* The Journal of Immunology, 2001. **166**(2): p. 747-753.
28. Mitchell, R.A. and R. Bucala, *Tumor growth-promoting properties of macrophage migration inhibitory factor (MIF).* Seminars in Cancer Biology, 2000. **10**(5): p. 359-366.
29. Hudson, J.D., et al., *A proinflammatory cytokine inhibits p53 tumor suppressor activity.* J Exp Med, 1999. **190**(10): p. 1375-82.
30. Akbar, S.M., et al., *Macrophage migration inhibitory factor in hepatocellular carcinoma and liver cirrhosis; relevance to pathogenesis.* Cancer Letters, 2001. **171**(2): p. 125-32.
31. Hira, E., et al., *Overexpression of macrophage migration inhibitory factor induces angiogenesis and deteriorates prognosis after radical resection for hepatocellular carcinoma.* Cancer, 2005. **103**(3): p. 588-598.
32. Zheng, Y., et al., *Hepatitis C virus non-structural protein NS4B can modulate an unfolded protein response.* Journal of Microbiology, 2005. **43**(6): p. 529-536.
33. Zhang, D.W., K.T. Jeang, and C.G.L. Lee, *p53 negatively regulates the expression of FAT10, a gene upregulated in various cancers.* Oncogene, 2006. **25**(16): p. 2318-2327.
34. Kao, C.F., et al., *Modulation of p53 transcription regulatory activity and post-translational modification by hepatitis C virus core protein.* Oncogene, 2004. **23**(14): p. 2472-83.
35. Tatrai, P., et al., *Agrin, a novel basement membrane component in human and rat liver, accumulates in cirrhosis and hepatocellular carcinoma.* Lab Invest, 2006. **86**(11): p. 1149-60.
36. Bataller, R. and D.A. Brenner, *Liver fibrosis.* The Journal of Clinical Investigation, 2005. **115**(2): p. 209-218.
37. Matsumoto, E., Y. Muragaki, and A. Ooshima, *Increased amount of serum type IV collagen peptide in human liver fibrosis as determined by enzyme-immunoassay with monoclonal antibodies.* Acta Pathol Jpn, 1989. **39**(4): p. 217-23.
38. Walsh, K.M., et al., *Basement membrane peptides as markers of liver disease in chronic hepatitis C.* Journal of hepatology, 2000. **32**(2): p. 325-330.
39. Hahn, E., et al., *Distribution of basement-membrane proteins in normal and fibrotic human-liver - collagen type-iv, laminin, and fibronectin.* Gut, 1980. **21**(1): p. 63-71.
40. Porter, S., et al., *The ADAMTS metalloproteinases.* Biochem J, 2005. **386**(Pt 1): p. 15-27.
41. Hall, N.G., et al., *ADAMTSL-3/punctin-2, a novel glycoprotein in extracellular matrix related to the ADAMTS family of metalloproteases.* Matrix Biology, 2003. **22**(6): p. 501-510.
42. Yang, W., et al., *Fatty acid synthase is up-regulated during hepatitis C virus infection and regulates hepatitis C virus entry and production.* Hepatology, 2008. **48**(5): p. 1396-1403.
43. Yamaguchi, A., et al., *Hepatitis C Virus Core Protein Modulates Fatty Acid Metabolism and Thereby Causes Lipid Accumulation in the Liver.* Digestive Diseases and Sciences, 2005. **50**(7): p. 1361-1371.
44. Kumari, M.V., M. Hiramatsu, and M. Ebadi, *Free radical scavenging actions of metallothionein isoforms I and II.* Free Radic Res, 1998. **29**(2): p. 93-101.
45. Carrera, G., et al., *Hepatic metallothionein in patients with chronic hepatitis C: relationship with severity of liver disease and response to treatment.* Am J Gastroenterol, 2003. **98**(5): p. 1142-9.

46. Kanda, M., et al., *Detection of metallothionein 1G as a methylated tumor suppressor gene in human hepatocellular carcinoma using a novel method of double combination array analysis.* International Journal of Oncology, 2009. **35**(3): p. 477-483.
47. Sakamoto, L.H., et al., *MT1G hypermethylation: a potential prognostic marker for hepatoblastoma.* Pediatr Res, 2010. **67**(4): p. 387-93.
48. Fernandes, A.P. and A. Holmgren, *Glutaredoxins: glutathione-dependent redox enzymes with functions far beyond a simple thioredoxin backup system.* Antioxid Redox Signal, 2004. **6**(1): p. 63-74.
49. Camaschella, C., et al., *The human counterpart of zebrafish shiraz shows sideroblastic-like microcytic anemia and iron overload.* Blood, 2007. **110**(4): p. 1353-1358.
50. Thursz, M., *Iron, haemochromatosis and thalassaemia as risk factors for fibrosis in hepatitis C virus infection.* Gut, 2007. **56**(5): p. 613-4.
51. Sheikh, M.Y., et al., *Hepatitis C virus infection: Molecular pathways to metabolic syndrome.* Hepatology, 2008. **47**(6): p. 2127-2133.
52. Aguado, B. and R.D. Campbell, *Characterization of a human lysophosphatidic acid acyltransferase that is encoded by a gene located in the class III region of the human major histocompatibility complex.* J Biol Chem, 1998. **273**(7): p. 4096-105.
53. Lee, H., E.J. Goetzl, and S. An, *Lysophosphatidic acid and sphingosine 1-phosphate stimulate endothelial cell wound healing.* American Journal of Physiology - Cell Physiology, 2000. **278**(3): p. C612-C618.
54. Watterson, K.R., et al., *Regulation of fibroblast functions by lysophospholipid mediators: Potential roles in wound healing.* Wound Repair and Regeneration, 2007. **15**(5): p. 607-616.
55. Hernandez-Gea, V. and S.L. Friedman, *Pathogenesis of Liver Fibrosis.* Annual Review of Pathology: Mechanisms of Disease, 2011. **6**(1): p. 425-456.
56. Ikeda, H., et al., *Effects of Lysophosphatidic Acid on Proliferation of Stellate Cells and Hepatocytes in Culture.* Biochemical and Biophysical Research Communications, 1998. **248**(2): p. 436-440.
57. Alison, M.R., S. Islam, and S. Lim, *Stem cells in liver regeneration, fibrosis and cancer: the good, the bad and the ugly.* J Pathol, 2009. **217**(2): p. 282-98.
58. Diez, J. and P. Iglesias, *The role of the novel adipocyte-derived hormone adiponectin in human disease.* Eur J Endocrinol, 2003. **148**(3): p. 293-300.
59. Arano, T., et al., *Serum level of adiponectin and the risk of liver cancer development in chronic hepatitis C patients.* International Journal of Cancer, 2011: p. n/a-n/a.
60. Park, J.E., et al., *Annexin A3 is a potential angiogenic mediator.* Biochemical and Biophysical Research Communications, 2005. **337**(4): p. 1283-1287.
61. Harashima, M., et al., *Annexin A3 Expression Increases in Hepatocytes and is Regulated by Hepatocyte Growth Factor in Rat Liver Regeneration.* Journal of Biochemistry, 2008. **143**(4): p. 537-545.
62. Wald, O., et al., *Involvement of the CXCL12/CXCR4 pathway in the advanced liver disease that is associated with hepatitis C virus or hepatitis B virus.* European Journal of Immunology, 2004. **34**(4): p. 1164-1174.
63. Wald, O., et al., *Chemokines in hepatitis C virus infection: pathogenesis, prognosis and therapeutics.* Cytokine, 2007. **39**(1): p. 50-62.
64. Komuves, L.G., et al., *Expression of epidermal growth factor and its receptor in cirrhotic liver disease.* J Histochem Cytochem, 2000. **48**(6): p. 821-30.
65. Ito, Y., et al., *Expression and clinical significance of erb-B receptor family in hepatocellular carcinoma.* Br J Cancer, 2001. **84**(10): p. 1377-83.

66. Berasain, C., et al., *The Epidermal Growth Factor Receptor: A Link Between Inflammation and Liver Cancer.* Experimental Biology and Medicine, 2009. **234**(7): p. 713-725.
67. Villanueva, A., et al., *Pivotal Role of mTOR Signaling in Hepatocellular Carcinoma.* Gastroenterology, 2008. **135**(6): p. 1972-1983.
68. Lee, D.F. and M.C. Hung, *All roads lead to mTOR - Integrating inflammation and tumor angiogenesis.* Cell Cycle, 2007. **6**(24): p. 3011-3014.
69. Marusawa, H., et al., *Hepatitis C virus core protein inhibits Fas- and tumor necrosis factor alpha-mediated apoptosis via NF-kappaB activation.* J Virol, 1999. **73**(6): p. 4713-20.
70. Saito, K., et al., *Hepatitis C virus core protein inhibits tumor necrosis factor alpha-mediated apoptosis by a protective effect involving cellular FLICE inhibitory protein.* J Virol, 2006. **80**(9): p. 4372-9.
71. Kawaguchi, T., et al., *Hepatitis C virus down-regulates insulin receptor substrates 1 and 2 through up-regulation of suppressor of cytokine signaling 3.* Am J Pathol, 2004. **165**(5): p. 1499-508.
72. Aytug, S., et al., *Impaired IRS-1/PI3-kinase signaling in patients with HCV: A mechanism for increased prevalence of type 2 diabetes.* Hepatology, 2003. **38**(6): p. 1384-1392.
73. Amitrano, L., et al., *Coagulation disorders in liver disease.* Semin Liver Dis, 2002. **22**(1): p. 83-96.
74. Yang, C., et al., *Integrin alpha1beta1 and alpha2beta1 are the key regulators of hepatocarcinoma cell invasion across the fibrotic matrix microenvironment.* Cancer Res, 2003. **63**(23): p. 8312-7.
75. Ito, Y., et al., *Activation of mitogen-activated protein kinases extracellular signal-regulated kinases in human hepatocellular carcinoma.* Hepatology, 1998. **27**(4): p. 951-958.
76. Soresi, M., et al., *Interleukin-6 and its soluble receptor in patients with liver cirrhosis and hepatocellular carcinoma.* World J Gastroenterol, 2006. **12**(16): p. 2563-8.
77. Naugler, W.E., et al., *Gender Disparity in Liver Cancer Due to Sex Differences in MyD88-Dependent IL-6 Production.* Science, 2007. **317**(5834): p. 121-124.
78. Ueki, T., et al., *Expression of hepatocyte growth factor and its receptor c-met proto-oncogene in hepatocellular carcinoma.* Hepatology, 1997. **25**(4): p. 862-866.
79. Sipula, I.J., N.F. Brown, and G. Perdomo, *Rapamycin-mediated inhibition of mammalian target of rapamycin in skeletal muscle cells reduces glucose utilization and increases fatty acid oxidation.* Metabolism, 2006. **55**(12): p. 1637-44.
80. Porstmann, T., et al., *SREBP Activity Is Regulated by mTORC1 and Contributes to Akt-Dependent Cell Growth.* Cell Metabolism, 2008. **8**(3): p. 224-236.
81. Ballardini, G., et al., *Hepatitis C virus (HCV) genotype, tissue HCV antigens, hepatocellular expression of HLA-A,B,C, and intercellular adhesion-1 molecules. Clues to pathogenesis of hepatocellular damage and response to interferon treatment in patients with chronic hepatitis C.* J Clin Invest, 1995. **95**(5): p. 2067-75.
82. Wraight, C.J., et al., *Human major histocompatibility complex class II invariant chain is expressed on the cell surface.* Journal of Biological Chemistry, 1990. **265**(10): p. 5787-5792.
83. Leng, L., et al., *MIF signal transduction initiated by binding to CD74.* J Exp Med, 2003. **197**(11): p. 1467-76.
84. Hosokawa, N., et al., *A novel ER alpha-mannosidase-like protein accelerates ER-associated degradation.* Embo Reports, 2001. **2**(5): p. 415-22.
85. Lee, C.G.L., et al., *Expression of the FAT 10 gene is highly upregulated in hepatocellular carcinoma and other gastrointestinal and gynecological cancers.* Oncogene, 2003. **22**(17): p. 2592-2603.
86. Ren, J.W., et al., *FAT10 plays a role in the regulation of chromosomal stability.* Journal of Biological Chemistry, 2006. **281**(16): p. 11413-11421.

87. Fukao, T., et al., *Enzymes of ketone body utilization in human tissues: Protein and messenger RNA levels of succinyl-coenzyme A (CoA):3-ketoacid CoA transferase and mitochondrial and cytosolic acetoacetyl-CoA thiolases.* Pediatric Research, 1997. **42**(4): p. 498-502.
88. Qin, X., A. Luke, and M.A. Lea, *dUTP pyrophosphatase and uracil-DNA glycosylase in rat liver and hepatomas.* International Journal of Biochemistry, 1992. **24**(3): p. 437-445.
89. Takatori, H., et al., *dUTP pyrophosphatase expression correlates with a poor prognosis in hepatocellular carcinoma.* Liver International, 2010. **30**(3): p. 438-446.
90. Okabe, M., et al., *Potential hepatic stem cells reside in EpCAM+ cells of normal and injured mouse liver.* Development, 2009. **136**(11): p. 1951-60.
91. Stokoe, D., et al., *MAPKAP kinase-2; a novel protein kinase activated by mitogen-activated protein kinase.* EMBO J, 1992. **11**(11): p. 3985-94.
92. Xu, L., S. Chen, and R.C. Bergan, *MAPKAPK2 and HSP27 are downstream effectors of p38 MAP kinase-mediated matrix metalloproteinase type 2 activation and cell invasion in human prostate cancer.* Oncogene, 2006. **25**(21): p. 2987-98.
93. Kumar, B., et al., *p38 Mitogen-Activated Protein Kinase–Driven MAPKAPK2 Regulates Invasion of Bladder Cancer by Modulation of MMP-2 and MMP-9 Activity.* Cancer Research, 2010. **70**(2): p. 832-841.
94. Yamauchi, T., et al., *Cloning of adiponectin receptors that mediate antidiabetic metabolic effects.* Nature, 2003. **423**(6941): p. 762-769.
95. Yamauchi, T., et al., *Targeted disruption of AdipoR1 and AdipoR2 causes abrogation of adiponectin binding and metabolic actions.* Nat Med, 2007. **13**(3): p. 332-339.
96. Moss, D., *Alkaline phosphatase isoenzymes.* Clin Chem, 1982. **28**(10): p. 2007-2016.
97. Ikura, Y., et al., *Expression of the hepatic endothelin system in human cirrhotic livers.* J Pathol, 2004. **204**(3): p. 304-10.
98. Yokomori, H., et al., *Enhanced Expression of Endothelin B Receptor at Protein and Gene Levels in Human Cirrhotic Liver.* The American Journal of Pathology, 2001. **159**(4): p. 1353-1362.
99. Hsu, L.S., et al., *Aberrant methylation of EDNRB and p16 genes in hepatocellular carcinoma (HCC) in Taiwan.* Oncol Rep, 2006. **15**(2): p. 507-11.
100. Breuhahn, K., T. Longerich, and P. Schirmacher, *Dysregulation of growth factor signaling in human hepatocellular carcinoma.* Oncogene, 2006. **25**(27): p. 3787-800.
101. Yang, J.D., I. Nakamura, and L.R. Roberts, *The tumor microenvironment in hepatocellular carcinoma: Current status and therapeutic targets.* Semin Cancer Biol, 2010.
102. Magis, A.T., et al., *Graphics processing unit implementations of relative expression analysis algorithms enable dramatic computational speedup.* Bioinformatics, 2011. **27**(6): p. 872-3.